Ann. Hum. Genet. (1998), 62, 159–179
Printed in Great Britain

Hidden linkage: a comparison of the affected sib pair (ASP) test and transmission/disequilibrium test (TDT)

R. E. McGINNIS
Department of Genetics, University of Pennsylvania School of Medicine, Philadelphia, PA 19104-6145

(Received 3.11.97. Accepted 16.3.98)

SUMMARY

I compare the transmission/disequilibrium test (TDT) and affected sib pair (ASP) test under a general algebraic model describing a bi-allelic disease locus. Assuming linkage to a bi-allelic marker, I derive two binomial probabilities, one for parental allele 'transmission' (P_t) which determines the magnitude of the TDT χ^2 statistic (χ^2_{tdt}), and a second for identity-by-descent (ibd) marker allele 'sharing' (P_s) which determines the magnitude of the ASP test statistic (χ^2_{asp}). I also consider the ASP test applied to a completely polymorphic marker and demonstrate that the probability of ASP marker allele sharing (P_s) is identical to P_s observed for a bi-allelic marker in equilibrium with the disease locus. I present a general framework for determining the power of the TDT and ASP test based on expressions for P_t, P_s and the proportion (H/F) of ascertained parents who are informative at the marker. Two previous analytic investigations of TDT power based on the work of Ott (1989), and Risch & Merikangas (1996) are shown to be special cases of this general framework. In addition, I show the relationship between the framework I present and a third analytic investigation of TDT power for multi-allelic markers based on the work of Sham & Curtis (1995).

[0325.02]

INTRODUCTION

Linkage has been demonstrated between insulin-dependent diabetes mellitus (IDDM) and the insulin gene region on chromosome 11p15.5 on the basis of linkage analysis by the transmission/disequilibrium test or TDT (McGinnis *et al.* 1991; Spielman *et al.* 1993). Linkage was demonstrated at the insulin 5′VNTR, a hypervariable marker that is extremely polymorphic, but whose VNTR alleles fall into two main size classes in Caucasians, thus forming a natural bi-allelic (+ / −) marker. The + alleles were discovered to be positively associated with IDDM in case-control studies (Bell *et al.* 1984). Subsequent studies then demonstrated linkage in families collected for Genetic Analysis Workshop 5 (GAW5) by TDT analysis of GAW5 parents who were heterozygous (+/−) under the 5′VNTR bi-allelic categories (Spielman *et al.* 1993; see also Thomson *et al.* 1989, Julier *et al.* 1991).

The very strong evidence for linkage provided by the TDT ($\chi^2 = 8.26$, $p < 0.005$) was both surprising and puzzling because identity-by-descent (ibd) sharing of 5′VNTR alleles in affected sib pairs (ASPs) yielded no evidence for linkage in the same GAW5 families. Indeed, evidence for linkage was completely undetected or 'hidden' because the proportion of alleles shared by ASPs did not exceed the null hypothesis value of 0.5 in two different types of ASP analysis. On one hand, there was no increase in ASP allele sharing when the analysis included all GAW5 families in which both parents were informative for any two lengths of 5′VNTR allele (Spielman *et al.* 1989; Cox & Spielman, 1989). On the other hand, when the analysis included only those ASP parents who were evaluated by the TDT, namely those heterozygous (+/−) when the 5′VNTR is con-

Address for correspondence: Dr. Ralph McGinnis, Senior Investigator, SmithKline Beecham, New Frontiers Science Park (North), Third Avenue, Harlow, Essex CM19 5AW.

one affected sib is shared (n_s) or 'unshared' (n_u) by the second affected sib; thus $n_s + n_u = n_{asp}$ is the sample size for χ^2_{asp} and equals the number of trios in the data set consisting of an informative parent and an ASP.

Unlike the ASP test which usually considers ASP allele sharing from parents informative for any two marker alleles, the TDT only considers parents heterozygous for two particular marker alleles (e.g. A/B only). For a set of nuclear families, the TDT counts the number of times each A/B parent transmitted allele A or B to individual affected offspring. As shown by Spielman *et al.* (1993), the χ^2 statistic for detecting linkage by the TDT is:

325.03

$$\chi^2_{tdt} = \frac{(n_a - n_b)^2}{(n_a + n_b)} = \frac{(n_a - n_b)^2}{n_{tdt}}$$

325.03

where n_a and n_b are the number of instances in which an A/B parent transmitted allele A or B, respectively, to an individual affected offspring; and thus $n_a + n_b = n_{tdt}$ is the sample size for χ^2_{tdt}.

Note that the algebraic expressions for χ^2_{asp} and χ^2_{tdt} are identical in form. In each χ^2, the denominator is the sample size of the data set. Thus, when sample size (n_{asp} or n_{tdt}) is fixed, the denominator is constant and the magnitude of each χ^2 is determined only by the size of the squared difference in the numerator [$(n_s - n_u)^2$ or $(n_a - n_b)^2$].

A key idea in this paper is that the magnitude of the numerator in each χ^2 is determined by a specific binomial probability. In the case of χ^2_{asp}, this is the probability of ASP 'allele *sharing*' or P_s, i.e. the probability that a randomly ascertained parent of an ASP transmitted the same marker allele (ibd) to both affected sibs. In the absence of linkage, $P_s = 0.5$. But when linkage is present $P_s > 0.5$, and the larger the value of $(P_s - 0.5)$, the more ASPs that exhibit allele sharing (n_s) and the higher the magnitude of χ^2_{asp}. Similarly, a second binomial probability denoted P_t (for probability of 'allele *transmission*') determines the size of χ^2_{tdt}. P_t is the probability that marker allele A was transmitted to a specific affected child by a randomly ascertained A/B parent of an ASP. When linkage and disequilibrium are present, $P_t \neq 0.5$ and the larger the value of $|P_t - 0.5|$, the greater the value of χ^2_{tdt}.

General algebraic model of linkage

At the beginning of Results, I give expressions for P_s and P_t based on the following general model: A bi-allelic marker with alleles A and B is linked to a bi-allelic disease locus with disease-predisposing allele D and non-predisposing allele d. The model allows any penetrance for the D/D, D/d and d/d genotypes (α, β, and γ, respectively) such that $1 \geq \alpha \geq 0$, $1 \geq \beta \geq 0$ and $1 \geq \gamma \geq 0$, and also assumes that no other locus underlies disease susceptibility. The recombination fraction (θ) between marker and disease locus is variable as are the population frequencies of the four marker-disease locus haplotypes [$f(AD) = c_1, f(Ad) = c_2, f(BD) = c_3, f(Bd) = c_4$, where $c_1 + c_2 + c_3 + c_4 = 1$].

325.04

Note that once the haplotype frequencies are specified, the population frequency (p) of disease allele D is known ($p = c_1 + c_3$), as are the frequencies ($m, 1-m$) of marker alleles A and B, respectively ($m = c_1 + c_2$; $1 - m = c_3 + c_4$). Furthermore, the coefficient of disequilibrium (δ) equals $c_1 c_4 - c_2 c_3$ and thus, when convenient, the haplotype frequencies can be expressed as $c_1 = mp + \delta$, $c_2 = m(1-p) - \delta$, $c_3 = (1-m)p - \delta$, and $c_4 = (1-m)(1-p) + \delta$.

325.04

RESULTS

Based on derivations in Appendix I, equations (1) and (2) show expressions for P_s and P_t in terms of standard genetic variables for the general bi-allelic model described above. Both expressions assume that parents are ascertained through a randomly selected ASP, and each expression applies

325.04 to an ascertained parent who is also heterozygous A/B at a bi-allelic marker. P_s is the probability that the parent transmitted the same marker allele to both affected sibs. P_t is the probability that the parent transmitted allele A to a particular affected child. 325.04

$$P_s = 0.5+(1-2\theta)^2\ \frac{c_1c_4+c_2c_3}{H}\quad p^2\frac{(\alpha-\beta)^2}{4}+2p(1-p)\frac{(\alpha-\gamma)^2}{16}+(1-p)^2\frac{(\beta-\gamma)^2}{4}$$

Equation 1

$$P_t = 0.5+(1-2\theta)\ \frac{c_1c_4-c_2c_3}{H}\quad p^2\ \frac{\alpha^2-\beta^2}{4}\ +2p(1-p)\ \frac{(\alpha+\beta)^2-(\beta+\gamma)^2}{16}\ +(1-p)^2\ \frac{\beta^2-\gamma^2}{4}$$

Equation 2

Note that the expressions for P_s and P_t are similar in form. When there is no linkage, both expressions equal 0.5; but when linkage is present an amount is added to 0.5 which, in each expression, is a product of three factors. In both expressions, the leftmost factor depends on the recombination fraction (θ), the middle factor on haplotype frequencies (c_1, c_2, c_3, c_4) and the quantity H (see Appendix I), and the rightmost factor on penetrances (α, β, γ) and the frequency (p) of disease allele D. Because they play analogous roles in each expression, I denote the leftmost factor in P_s and P_t as L_s and L_t, respectively; and similarly denote the rightmost factor as R_s and R_t, and middle factor as M_s and M_t. Thus, $P_s = 0.5+L_sM_sR_s$ while $P_t = 0.5+L_tM_tR_t$.

Why $|P_t-0{\cdot}5| > (P_s-0{\cdot}5)$ when disequilibrium is extreme

As described in the Introduction, the ASP approach failed to detect linkage at the insulin 5′VNTR because the proportion of marker allele sharing in ASPs was close to 0.5, i.e. $P_s \approx 0.5$ whether the 5′VNTR was treated as bi-allelic or as highly polymorphic. By contrast, the TDT was able to detect linkage in the same families because $P_t \geqslant 0.60$ (see Spielman *et al.* 1993). Thus, regardless of differences in relative sample size (n_{asp}, n_{tdt}) for χ^2_{asp} and χ^2_{tdt}, the relative magnitudes of $(P_s-0.5)$ and $|P_t-0.5|$ are often the critical factor that causes a substantial difference in power for χ^2_{asp} and χ^2_{tdt}.

It is therefore interesting that analysis of equations (1) and (2) (see below) shows that when disequilibrium (δ) reaches its most positive value (δ_{max}) or its most negative value (δ_{min}), the magnitudes of P_s and P_t are such that: (a) $|P_t-0.5|$ and $(P_s-0.5)$ are both maximized and (b) $|P_t-0.5| > (P_s-0.5)$. This dependence on δ of $|P_t-0.5|$ and $(P_s-0.5)$ also has other important implications since the value of P_s for a *completely polymorphic* marker is identical to the P_s value of a bi-allelic marker in *equilibrium* ($\delta = 0$) with a bi-allelic disease locus (see Appendix IV). Therefore, if ASP allele sharing for a completely polymorphic marker is denoted by $(P_s-0.5)_{\delta=0}$, then P_t and P_s for any bi-allelic marker in extreme disequilibrium with the bi-allelic disease locus are such that: $|P_t-0.5| > (P_s-0.5) > (P_s-0.5)_{\delta=0}$.

To understand the pivotal role of δ in maximizing $|P_t-0.5|$ and $(P_s-0.5)$, and in determining their relative magnitudes, consider the three corresponding factors in P_t and P_s. By inspection, $L_t = (1-2\theta) \geqslant L_s = (1-2\theta)^2$, and when disequilibrium is present, θ should be near 0 and hence $L_t \approx L_s \approx 1$, the maximum value of each factor. Furthermore, as shown below, R_t is substantially greater than R_s, and the difference between the two factors is independent δ and θ since R_t and R_s depend only on the properties of the disease locus (p, α, β, γ). Therefore, $L_t \geqslant L_s$ and $R_t > R_s$, so it follows that $|P_t-0.5|$ would always exceed $(P_s-0.5)$ were it not for the influence of the remaining two factors in equations (1) and (2) (M_t and M_s).

Note, then, that M_t and M_s have the same denominator (H as defined in Appendix I); but the numerator of M_t is $\delta = c_1c_4-c_2c_3$, while the numerator of M_s is the two components of δ *added*

together ($c_1 c_4 + c_2 c_3$) implying that $M_s \geq |M_t|$. Since $|M_t|$ reaches its minimum value of 0 at equilibrium ($\delta = 0$), while M_s is always positive, it follows that $(P_s - 0.5) > |P_t - 0.5| \approx 0$ in an interval of δ values around $\delta = 0$. However, in Appendix III, I assume that marker allele frequency (m) and disease allele frequency (p) are fixed, and then show that $|M_t| = M_s$ at $\delta = \delta_{max}$ and at $\delta = \delta_{min}$. I also show that $|M_t|$ and M_s are both maximized at one of the two extreme δ values (δ_{max} *or* δ_{min}). Therefore, for any bi-allelic marker (i.e. any m and p), when δ equals δ_{max} or δ_{min}, $|P_t - 0.5|$ and $(P_s - 0.5)$ are maximized since $|M_t|$ and M_s are maximized; furthermore, because $|M_t| = M_s$ and $L_t \approx L_s$, the greater magnitude of R_t compared to R_s drives $|P_t - 0.5|$ higher than $(P_s - 0.5)$.

Why $R_t > R_s$

To understand why $R_t > R_s$, note first that both factors contain three components that are multiplied by the coefficients $p^2/4$, $[2p(1-p)]/16$ and $(1-p)^2/4$, respectively. In R_s, each component has the form $(U-V)^2$ while the corresponding component in R_t is $(U^2 - V^2)$ where U and V (for the components multiplied by $p^2/4$, $[2p(1-p)]/16$ and $(1-p)^2/4$) are, respectively, $U = \alpha, \alpha+\beta, \beta$ and $V = \beta, \beta+\gamma, \gamma$. Under the assumption that D/D penetrance exceeds d/d penetrance ($\alpha > \gamma$) and that D/d penetrance (β) lies somewhere between ($\alpha \geq \beta \geq \gamma$), each component in R_t $[U^2 - V^2 = (U+V)(U-V)]$ must exceed its counterpart in R_s $[(U-V)^2]$ since $(U+V) > (U-V)$. The only exceptions occur when mode of inheritance is dominant ($\alpha = \beta$) or recessive ($\beta = \gamma$) in which case one pair of analogous components in R_t and R_s are equal; however, the other two components in R_t still exceed their counterparts in R_s, and thus $R_t > R_s$.

To assess how the elevation of R_t above R_s is influenced by the degree of risk conferred by the disease locus, the risk can be quantified by considering the penetrance of the D/D homozygote (α) to be r times greater than the penetrance of the d/d homozygote (γ). Thus, $\alpha = r\gamma$ and the penetrance of D/d (β) can be considered to fall between α and γ by letting $\beta = \gamma + x(\alpha - \gamma) = \gamma + x(r-1)\gamma$ where x is a number between 0 and 1. Based on this parameterization, the $\alpha:\gamma$ penetrance ratio (r) can be evaluated for its influence on R_t and R_s by dividing each component in R_t by its conterpart in R_s which yields the ratios: [0325.05]

$$\frac{1}{1-x} \quad 1 + x + \frac{2}{r-1} \quad , \quad 2 \quad x + \frac{1}{2} + \frac{2}{r-1} \quad , \quad 1 + \frac{1}{x} \quad \frac{2}{r-1} \quad .$$

Note that r appears in each ratio only in the term $2/(r-1)$ implying that R_t/R_s increases monotonically as r decreases. Thus, the elevation of R_t above R_s is most extreme for susceptibility loci causing a modest increase in disease risk as indicated by low values of r. [0325.05]

In Tables 1 and 2 below, I show values of P_s and P_t when $r = 2$ and $r = 4$, respectively. In these tables, $(P_s - 0.5) \approx 0$ indicating that linkage would be difficult to detect by the ASP approach; but $|P_t - 0.5|$ is much greater than $(P_s - 0.5)$ when disequilibrium is extreme, thus illustrating that at low r values, R_t drives $|P_t - 0.5|$ to levels that provide strong evidence for linkage.

Power of χ^2_{asp} *and* χ^2_{tdt}

I now show how to calculate and compare the power of χ^2_{asp} and χ^2_{tdt} when (a) both tests are applied to a bi-allelic marker or (b) the TDT is applied to a bi-allelic marker but the ASP test considers a completely polymorphic marker. I assume the two tests evaluate a series of S randomly ascertained parents of one or more ASPs, and that each test considers one ASP per parent. In the Discussion, I explain how to calculate the proportion (H/F) of the S parents who are informative at a bi-allelic marker. (The quantity F is proportional to the population frequency of parents who have two or more affected children, and H is proportional to the frequency of such parents who are also heterozygous at the marker.) Thus H/F determines the sample size for χ^2_{asp} and χ^2_{tdt} (see [0325.06]

Proportion of A/B parents in ascertained families). For instance, if both tests evaluate the same bi-allelic marker, then sample size for χ^2_{asp} is $n_{asp} = (H/F)S$, while sample size for χ^2_{tdt} is twice as large ($n_{tdt} = 2(H/F)S$) since χ^2_{asp} counts pairs of transmitted alleles while χ^2_{tdt} counts individual alleles. Based on these sample sizes (n_{asp}, n_{tdt}) and the values of P_s and P_t, the power of χ^2_{asp} and χ^2_{tdt} are determined from the binomial distributions

$$\frac{n_{asp}!}{n_a!n_u!}P_s^{n_a}(1-P_s)^{n_u} \quad \text{and} \quad \frac{n_{tdt}!}{n_a!n_b!}P_t^{n_a}(1-P_t)^{n_b},$$

respectively, as explained below.

Similarly, the power of χ^2_{asp} when applied to a completely polymorphic marker can also be determined from the appropriate binomial distribution, but $n_{asp} = S$ since all parents are informative, and $P_s = 0.5+(1-2\theta)^2[p(1-p)/F]R_s$ as shown in Appendix IV. Interestingly, this expression for P_s when the marker is completely polymorphic is identical to P_s for a bi-allelic marker in equilibrium with a bi-allelic disease locus. This can be verified by setting $\delta = 0$ in $c_1 = mp+\delta$, $c_2 = m(1-p)-\delta$, $c_3 = (1-m)p-\delta$, $c_4 = (1-m)(1-p)+\delta$ and substituting for the four haplotype frequencies in the expression for H (see Appendix I) and in equation (1).

Based on sample size (n_{asp}, n_{tdt}) and binomial probability (P_s, P_t), two binomial distributions are generated which can be used to calculate the power of χ^2_{asp} and χ^2_{tdt} as described in Appendix II. Specifically, the power of χ^2_{asp} or the probability that $\chi^2_{asp} > L$ (a significance cutpoint) is equal to the portion of the binomial distribution

$$\frac{n_{asp}!}{n_a!n_u!}P_s^{n_a}(1-P_s)^{n_u} \quad \text{for which} \quad n_a > \frac{n_{asp}}{2}+\frac{\sqrt{(n_{asp}L)}}{2}.$$

Similarly, if marker allele A is associated with disease, the power of χ^2_{tdt} is estimated by the portion of the binomial distribution

$$\frac{n_{tdt}!}{n_a!n_b!}P_t^{n_a}(1-P_t)^{n_b} \quad \text{for which} \quad n_a > \frac{n_{tdt}}{2}+\frac{\sqrt{(n_{tdt}L)}}{2}.$$

Thus, standard tables giving the normal approximation to the binomial distribution (Pearson & Hartley, 1954; Weir, 1996) provide precise power values for virtually any sample size (n_{asp}, n_{tdt}), binomial probability (P_s, P_t), and significance level.

Comparison of TDT and ASP power

Here I illustrate how the equations for P_t, P_s and H/F can be used to compare the power of χ^2_{tdt} and χ^2_{asp}. I assume the two tests consider markers that are tightly linked ($\theta = 0$) to bi-allelic disease loci with additive mode of inheritance ($\beta = (\alpha+\gamma)/2$) and for which the $\alpha:\gamma$ penetrance ratio is $r = 2$, $r = 4$ or $r = 10$. Penetrance ratios of $r = 2$, 4 and 10 were chosen as being somewhat representative of the entire genetic parameter space since I have found that P_t and P_s increase rapidly as r increases from 2 to 6 with smaller, asymptotic increases in P_t and P_s for $r > 10$. Furthermore, additive mode of inheritance may also be regarded as being somewhat representative since results from other modes of inheritance do not, in general, substantially differ from results presented here. In the tables below, I compare χ^2_{tdt} and χ^2_{asp} when both tests consider the same bi-allelic marker, or when χ^2_{asp} considers a fully informative marker and χ^2_{tdt} evaluates a nearby bi-allelic marker. Such single test comparisons would be occasioned by: (a) TDT and ASP analysis of a marker that gave 'suggestive' evidence of linkage and disease-association in other families or in comparisons of allele frequencies in cases and unrelated controls; or (b) TDT and ASP analysis of markers near a candidate gene suspected of increasing disease susceptibility.

Table 1. *ASP and TDT power for $\alpha:\gamma$ penetrance ratio of $r = 2$*[a]

	ASP test of fully informative marker[c]			TDT and ASP test of the same bi-allelic marker									
				$\delta = \delta_{max}$[b]					$\delta = \frac{1}{2}\delta_{max}$[b]				
							Power					Power	
dis. allele freq(p)[b]	P_s	H/F	Power ASP	mkr. allele freq(m)[b]	P_t	P_s	H/F	TDT	ASP	P_t	P_s	H/F	TDT	ASP
$p = 0.60$	0.506	1.0	0.08	$m = 0.75$	0.568	0.511	0.33	0.60	0.08	0.532	0.507	0.35	0.19	0.07
				$m = 0.50$	0.560	0.510	0.50	0.67	0.09	0.530	0.507	0.50	0.22	0.07
				$m = 0.25$	0.537	0.506	0.40	0.26	0.07	0.519	0.506	0.39	0.10	0.07
$p = 0.40$	0.507	1.0	0.09	$m = 0.75$	0.553	0.509	0.34	0.42	0.07	0.525	0.508	0.36	0.13	0.07
				$m = 0.50$	0.573	0.513	0.50	0.83	0.10	0.536	0.509	0.50	0.30	0.08
				$m = 0.25$	0.563	0.511	0.43	0.65	0.09	0.534	0.508	0.40	0.23	0.07
$p = 0.15$	0.506	1.0	0.08	$m = 0.75$	0.526	0.505	0.36	0.14	0.06	0.513	0.505	0.37	0.07	0.07
				$m = 0.50$	0.537	0.507	0.50	0.32	0.07	0.518	0.506	0.50	0.11	0.07
				$m = 0.25$	0.585	0.512	0.42	0.67	0.09	0.534	0.509	0.40	0.23	0.08

[a] ASP power (1-tailed test) and TDT power (2-tailed test) for a significance level of 0.05 and sample size of 200 families; thus $n_{asp} = 400\ H/F$ and $n_{tdt} = 800\ H/F$.
[b] δ_{max} (δ_{min}) is most positive (most negative) value of disequilibrium for bi-allelic marker and disease locus with allele frequencies m and p, respectively; power results shown for δ_{max} ($1/2\delta_{max}$) at $m = 0.75$, 0.5 and 0.25 equal power results for δ_{min} ($1/2\delta_{min}$) when $m = 0.25$, 0.5 and 0.75, respectively.
[c] P_s for a fully informative marker is identical to P_s for a bi-allelic marker at $\delta = 0$.

Table 2. *ASP and TDT power for $\alpha:\gamma$ penetrance ratio of $r = 4$*[a]

	ASP test of fully informative marker[c]			TDT and ASP test of the same bi-allelic marker									
				$\delta = \delta_{max}$[b]					$\delta = \frac{1}{2}\delta_{max}$[b]				
							Power					Power	
dis. allele freq(p)[b]	P_s	H/F	Power ASP	mkr. allele freq(m)[b]	P_t	P_s	H/F	TDT	ASP	P_t	P_s	H/F	TDT	ASP
$p = 0.60$	0.516	1.0	0.16	$m = 0.75$	0.620	0.534	0.30	0.96	0.18	0.553	0.522	0.34	0.42	0.13
				$m = 0.50$	0.597	0.527	0.40	0.97	0.19	0.548	0.519	0.50	0.52	0.13
				$m = 0.25$	0.556	0.516	0.42	0.54	0.11	0.530	0.515	0.40	0.19	0.10
$p = 0.40$	0.525	1.0	0.26	$m = 0.75$	0.607	0.534	0.31	0.93	0.19	0.548	0.527	0.34	0.35	0.15
				$m = 0.50$	0.636	0.543	0.48	0.99	0.32	0.566	0.530	0.50	0.75	0.21
				$m = 0.25$	0.607	0.534	0.46	0.99	0.24	0.568	0.527	0.42	0.57	0.17
$p = 0.15$	0.530	1.0	0.33	$m = 0.75$	0.570	0.526	0.33	0.63	0.15	0.533	0.528	0.35	0.19	0.16
				$m = 0.50$	0.594	0.536	0.49	0.96	0.26	0.547	0.531	0.50	0.53	0.22
				$m = 0.25$	0.644	0.555	0.48	0.99	0.45	0.581	0.540	0.43	0.88	0.27

[a] ASP power (1-tailed test) and TDT power (2-tailed test) for a significance level of 0.05 and sample size of 200 families; thus $n_{asp} = 400\ H/F$ and $n_{tdt} = 800\ H/F$.
[b] δ_{max} (δ_{min}) is most positive (most negative) value of disequilibrium for bi-allelic marker and disease locus with allele frequencies m and p, respectively; power results shown for δ_{max} ($1/2\delta_{max}$) at $m = 0.75$, 0.5 and 0.25 equal power results for δ_{min} ($1/2\delta_{min}$) when $m = 0.25$, 0.5 and 0.75, respectively.
[c] P_s for a fully informative marker is identical to P_s for a bi-allelic marker at $\delta = 0$.

Comparison of the TDT and ASP test

165

Paragraph [0325.10]

Table 3. *ASP and TDT power for α:γ penetrance ratio of $r = 10^a$*

	ASP test of fully informative marker[c]			TDT and ASP test of the same bi-allelic marker									
				$\delta = \delta_{max}$[b]					$\delta = \frac{1}{2}\delta_{max}$[b]				
							Power					Power	
dis. allele freq(p)[b]	P_a	H/F	Power ASP	mkr. allele freq(m)[b]	P_t	P_a	H/F	TDT	ASP	P_t	P_a	H/F	TDT	ASP
$p = 0.60$	0.527	1.0	0.28	$m = 0.75$	0.660	0.560	0.27	0.99	0.34	0.567	0.537	0.33	0.59	0.21
				$m = 0.50$	0.621	0.546	0.48	0.99	0.36	0.559	0.531	0.50	0.66	0.22
				$m = 0.25$	0.568	0.526	0.43	0.72	0.17	0.536	0.525	0.40	0.25	0.16
$p = 0.40$	0.547	1.0	0.59	$m = 0.75$	0.654	0.569	0.28	0.99	0.42	0.566	0.552	0.33	0.56	0.33
				$m = 0.50$	0.686	0.583	0.47	0.99	0.74	0.589	0.555	0.49	0.94	0.46
				$m = 0.25$	0.636	0.560	0.48	0.99	0.51	0.576	0.549	0.43	0.81	0.36
$p = 0.15$	0.580	1.0	0.94	$m = 0.75$	0.633	0.577	0.31	0.99	0.53	0.560	0.577	0.34	0.51	0.56
				$m = 0.50$	0.668	0.596	0.49	0.99	0.86	0.582	0.584	0.50	0.91	0.77
				$m = 0.25$	0.727	0.630	0.54	0.99	0.99	0.632	0.599	0.47	0.99	0.86

[a] ASP power (1-tailed test) and TDT power (2-tailed test) for a significance level of 0.05 and sample size of 200 families; thus $n_{asp} = 400\ H/F$ and $n_{tdt} = 800\ H/F$.
[b] δ_{max} (δ_{min}) is most positive (most negative) value of disequilibrium for bi-allelic marker and disease locus with allele frequencies m and p, respectively; power results shown for δ_{max} ($1/2\delta_{max}$) at $m = 0.75$, 0.5 and 0.25 equal power results for δ_{min} ($1/2\delta_{min}$) when $m = 0.25$, 0.5 and 0.75, respectively.
[c] P_a for a fully informative marker is identical to P_a for a bi-allelic marker at $\delta = 0$.

Comparison of the TDT and ASP test

Paragraph [0325.10]

In Table 1 ($r = 2$), Table 2 ($r = 4$) and Table 3 ($r = 10$), column 1 shows disease allele frequency (p) for a bi-allelic disease locus, and columns 2, 3 and 4 show results for χ^2_{asp} applied to a fully informative marker. The remaining columns in each table list χ^2_{tdt} and χ^2_{asp} results for a linked bi-allelic marker whose allele frequency (m) is listed in column 5. Results are given for each value of m (0.75, 0.50, 0.25) assuming positive disequilibrium between the bi-allelic marker allele and disease allele is maximal ($\delta = \delta_{max}$) or half-maximal ($\delta = \frac{1}{2}\delta_{max}$) where $\delta_{max} = \min[(1-m)p, (1-p)m]$. TDT power (two-tailed test) and ASP power (one-tailed test) are for a significance level of 0.05 and are based on a sample size of 200 families (i.e. 400 parent-ASP trios) and thus $n_{asp} = 400(H/F)$ while $n_{tdt} = 800(H/F)$.

The ASP test can detect linkage over long distances ($\theta \geqslant 0$) and in the absence of disequilibrium ($\delta = 0$); but the TDT has no power when $\delta = 0$ and hence can detect linkage only over short distances (generally less than 1 cM). Yet when disequilibrium is half-maximal or greater ($\delta \geqslant \frac{1}{2}\delta_{max}$), Tables 1, 2 and 3 each show that TDT power almost always exceeds ASP power whether χ^2_{asp} is applied to a fully informative or bi-allelic marker. When $r = 2$ (Table 1), linkage is virtually undetectable by χ^2_{asp} since $(P_s - 0.5) \leqslant 0.13$ and ASP power is 0.10 or lower; by contrast, the TDT is able to detect linkage but TDT power exceeds 0.50 only when δ is close to δ_{max} and allele frequencies (m, p) are similar in magnitude at the marker and disease locus. For $r = 4$ (Table 2), ASP power is increased but still relatively low ($\leqslant 0.33$) for fully informative markers and for most bi-allelic markers. TDT power is also substantially higher and, for most markers, exceeds 0.95 when $\delta = \delta_{max}$ and exceeds 0.50 when $\delta = \frac{1}{2}\delta_{max}$, thus indicating that when $r = 4$, the TDT could demonstrate linkage to many disease loci whose linkage might be difficult or impossible to establish by the ASP test.

For $r = 10$ (Table 3), TDT power is reasonably high ($\geqslant 0.66$) when $\delta \geqslant \frac{1}{2}\delta_{max}$ and ASP power is also elevated (> 0.50) except at the highest disease allele frequency shown ($p = 0.6$) where ASP power is 0.28 for a fully informative marker. Thus as r increases from 2 to 10, the tables show that P_s and ASP power increase substantially and hence, when $r = 10$, the relative power advantage of the TDT is diminished. Nevertheless, as indicated by lower ASP power when $p = 0.6$ at $r = 10$ (Table 3), ASP power at elevated disease allele frequencies ($p > 0.6$) remains low (< 0.50) even when $r \to \infty$ (data not shown; table available from the author). For example, if the same power analysis shown in Tables 1–3 were conducted for $r = \infty$ (i.e. $\gamma = 0$) then for a fully informative marker and disease allele frequency of $p = 0.75$, P_s and ASP power would be 0.519 and 0.19, respectively. By contrast, TDT power would be much higher ($\geqslant 0.80$) but only when $\delta \geqslant \frac{3}{4}\delta_{max}$ and m is close to $p = 0.75$ (i.e. $0.65 \leqslant m \leqslant 0.85$).

In concluding this section, I emphasize that Tables 1–3 show that when the disease locus and marker are bi-allelic, TDT power is substantially increased if the disease allele and positively associated marker allele have similar frequencies. Müller-Myshok & Abel (1997) independently made a similar observation, but they emphasized the weakness of TDT power when the m/p ratio departs from unity and δ is not close to δ_{max}. However, the tables illustrate that similar frequencies for the disease allele and associated marker allele can increase TDT power to reasonably high levels even when the m/p ratio substantially differs from 1 and δ is much lower than δ_{max}. For example, in Table 3 ($r = 4$), note that when $\delta = \frac{1}{2}\delta_{max}$ and $p = 0.15$, a similar frequency ($m = 0.25$) for the disease-associated marker allele produces TDT power of 0.86 and P_t of 0.581; but when $p = 0.15$ and $m = 0.5$ at $\delta = \frac{1}{2}\delta_{max}$, TDT power and P_t fall to 0.53 and 0.547, respectively. The difference in TDT power for these two situations can also be quantified by calculating the mean value of χ^2_{tdt} based on a sample of 200 ASP families and the values of P_t and H/F in Table 4 [i.e. $\chi^2_{tdt} = 800(H/F)(2P_t - 1)^2$]. When $p = 0.15$ and $m = 0.5$, $\chi^2_{tdt} = 3.53$ yielding a significance level of $p = 0.06$; but when $p = 0.15$ and $m = 0.25$, $\chi^2_{tdt} = 9.02$ for a significance level of $p < 0.003$. The large

[0325.1]

difference in significance level (0.06 versus 0.003) and power (0.53 versus 0.86) illustrated by this example indicates that careful attention to allele frequencies at bi-allelic markers may play an important role in future efforts to map susceptibility loci.

DISCUSSION

The equations for P_s, P_t and H/F enable comparison of TDT and ASP power for the same family data, since the three expressions assume random ascertainment of parents of *two* or more affected children. However, the TDT can be applied to families with a single affected child, so in Appendix I, I derive an expression analogous to P_t (denoted P_t^*) which gives the probability that allele A was transmitted to an affected child by a randomly ascertained A/B parent of *one* or more affected offspring. The derivation of P_t^* is almost identical to that of P_t, and hence the algebraic form of P_t^* is similar to that of P_t and P_s:

325.12

$$P_t^* = 0.5 + (1-2\theta)\,\frac{c_1c_4 - c_2c_3}{H^*}\quad p^2\frac{(\alpha-\beta)}{2} + p(1-p)\frac{(\alpha-\gamma)}{2} + (1-p)^2\frac{(\beta-\gamma)}{2}\ ,$$

Equation 3

where $$H^* = 2(c_1c_4 + c_2c_3)\frac{p\alpha - p\gamma + \beta + \gamma}{2} + 2c_1c_3\{p\alpha - p\beta + \beta\} + 2c_2c_4\{p\beta - p\gamma + \gamma\}$$

325.12

Previous analyses of TDT power are special cases of the current analysis

I now show that two previous analytic investigations of TDT power (Terwilliger & Ott, 1992; Risch & Merikangas, 1996) are special cases of the current analysis. These two analytic investigations and a third by Sham & Curtis (1995) as well as simulation and computer-based analyses of TDT power (Schaid & Sommer, 1994; Clerget-Darpoux *et al.* 1995; Kaplan *et al.* 1997) all assumed the disease locus to be bi-allelic. Sham & Curtis (1995) and Kaplan *et al.* (1997) considered a multi-allele marker locus, but the other analyses assumed either a bi-allelic marker or a direct test of the disease polymorphism itself, and each analysis examined TDT power for one or several specific modes of inheritance.

Terwilliger & Ott (1992) considered a recessive disease with no phenocopies in families ascertained through a single affected child. By investigating the same recessive model, Ott (1989) had previously derived an algebraic probability for transmission of each marker allele (denoted H and h) to affected offspring by heterozygous H/h parents and by both types of homozygous parent (see Ott's table II). Thus, power results for the TDT (McNemar's test in figure 3 of Terwilliger & Ott) can be derived by using Ott's table II to compute P_t^* for the recessive, zero-phenocopy model. P_t^* is derived by considering only the two probabilities in Ott's table for heterozygous H/h parents, and by dividing the transmission probability for allele H by the sum of the probabilities for allele H and allele h to yield:

$$P_t^* = \frac{(m+\delta/p)(1-m) - \theta\delta/p}{(m+\delta/p)(1-m) + m[(1-m) - \delta/p]},$$

where, substituting my notation for Ott's, m is the frequency of marker allele H (or, alternatively, my allele A), p is the frequency of disease allele D, and δ is the coefficient of disequilibrium. This expression for P_t^* is seen to be a special case of equation (3) by making the appropriate penetrance substitutions ($\alpha > 0, \beta = \gamma = 0$) into my expressions for P_t^* and H^*, and by expressing c_1, c_2, c_3 and c_4 in terms of m, p and δ according to standard expressions given above (see General algebraic model of linkage).

Risch & Merikangas (1996) compared TDT and ASP power for an intermediate mode of inheritance in which D/d penetrance (β) is a multiple (k) of d/d penetrance and $r = k^2$. For their analysis,

the TDT was assumed to test the disease locus itself or a perfectly associated bi-allelic marker and, under this assumption, Risch & Merikangas found that $P_t^* = P_t = k/(1+k)$. With appropriate substitutions ($\theta = 0, \alpha = k^2\gamma, \beta = k\gamma, c_1 = p, c_4 = 1-p,\ c_2 = c_3 = 0$), equations (2) and (3) for P_t and P_t^* also simplify to $k/(1+k)$, thus agreeing that for this particular model, P_t and P_t^* are (a) identical and (b) independent of disease allele frequency. For many other genetic models, P_t and P_t^* do appear to be similar in value (though not identical). However, graphical analysis of P_t shows that the value of P_t is independent of disease allele frequency only for the particular mode of inheritance considered by Risch & Merikangas (graphs showing this are available from the author).

The expression in Risch & Merikangas (1996) for P_s (their Y) when a marker is fully informative can also be shown to be a special case of equation (1) for P_s. This is verified by substituting the appropriate mode of inheritance parameters ($\alpha = k^2\gamma, \beta = k\gamma$) into equation (1) and also by substituting parameters for a closely linked marker in equilibrium with the disease locus (i.e. $\theta = 0$ and $\delta = 0$ in $c_1 = mp+\delta$, $c_2 = m(1-p)-\delta$, $c_3 = (1-m)p-\delta$ and $c_4 = (1-m)(1-p)+\delta$). In Appendix IV and Results (see Power of χ^2_{ilsp} and χ^2_{ild}), I showed that when a bi-allelic marker is in equilibrium with the disease locus, equation (1) for P_s is identical to the expression for P_s when a marker is fully informative as was assumed by Risch & Merikangas (1996).

Proportion of A/B parents in ascertained families

Since power analyses often calculate power for a specific number of ascertained families, the proportion of informative A/B parents in such families must be calculated to determine the subset of parents to which the TDT or ASP test is applied when a marker is bi-allelic. Among parents ascertained through one affected child, it can be shown that the expected proportion of A/B parents is H^*/F^* where H^* is as previously defined (see equation (3)) and $F^* = p^2\alpha+2p(1-p)\beta+(1-p)^2\gamma$. Similarly, the expected proportion of A/B parents among those ascertained through an ASP can be shown to be H/F where H is as defined in Appendix I and

$$F = p^4\alpha^2+4p^3(1-p)\left(\frac{\alpha+\beta}{2}\right)^2+2p^2(1-p)^2\beta^2+4p^2(1-p)^2\left(\frac{\alpha+2\beta+\gamma}{4}\right)^2+4p(1-p)^3\left(\frac{\beta+\gamma}{2}\right)^2+(1-p)^4\gamma^2.$$

With appropriate substitutions ($\alpha = k^2\gamma, \beta = k\gamma$), the expressions H^*/F^* and H/F reduce to the corresponding expressions given by Risch & Merikangas (1996) for the proportion of heterozygous parents found in families having at least one and two affected children, respectively. Furthermore, for the recessive, zero-phenocopy disease considered by Ott (1989), the sum of the two probabilities for heterozygous parents in Ott's table II gives the proportion of heterozygous parents in families ascertained through a single affected child. When appropriate substitutions are made ($\alpha > 0, \beta = \gamma = 0$), the expression H^*/F^* also reduces to the proportion of heterozygous parents predicted by Ott's table. It is also important to note that Sham & Curtis (1995) derived a table of probabilities analogous to Ott's table II, except that their table 3 has entries for a variable number of marker alleles and their probabilities describe a general model of disease. If table 3 of Sham & Curtis is assumed to have only two marker alleles, then the two probabilities for heterozygous parents predict a probability of allele transmission identical to P_t^* (equation (3)) as well as a proportion of heterozygous parents in ascertained families which is identical to H^*/F^*.

Power of χ^2_{tdt} for a multi-allelic marker

So far the four haplotype frequencies (c_1, c_2, c_3, c_4) have represented a bi-allelic marker linked to a bi-allelic disease locus; but these frequencies could also correspond to any two marker alleles (a_i, a_j) of a multi-allelic marker linked to a bi-allelic disease locus [$c_1 = f(a_i D), c_2 = f(a_i d), c_3 =$

$f(a_j D), c_4 = f(a_j d)$]. The expression for H^*/F^* (or H/F) would then be the proportion of ascertained parents expected to be heterozygous for a_i and a_j, and the expression for P_t^* (or P_t) would be the conditional probability that a_i/a_j parents transmit allele a_i to affected offspring. Thus, in principle, the expressions for P_t^* (or P_t) and H^*/F^* (or H/F) could be used to investigate the power of any strategy for applying the TDT to a multi-allelic marker.

Here I briefly discuss a strategy recommended by Spielman & Ewens (1996) in which χ^2_{tdt} is calculated for each allele i of a multi-allelic marker ($i = 1$ to k) by evaluating parents heterozygous for allele i and the other alleles grouped together (non-i). The marker is then tested for linkage by evaluating the statistical significance of the largest of the k χ^2_{tdt}'s using significance cutpoints adjusted for multiple testing and non-independence of the chi-squares (see Ewens & Spielman [1997] for a table of these cutpoints). The power of this procedure can be estimated for any multi-allelic marker model as follows: For each i/non-i determine the haplotype frequencies c_1, c_2, c_3, c_4 and calculate the associated values of P_t^* and H^*/F^* (or P_t, H/F and P_s). Then determine the i/non-i likely to give the highest χ^2_{tdt} by calculating the expected value of each χ^2_{tdt} [$E(\chi^2_{tdt})$]. (For S parents of an ASP, it can be shown that $E(\chi^2_{tdt}) = 2(S-1)(2P_t-1)^2+2P_s$ while for singletons, $E(\chi^2_{tdt}) = (S-1)(2P_t^*-1)^2+1$.) For the i/non-i giving the highest $E(\chi^2_{tdt})$, TDT power would then be determined exactly as for a bi-allelic marker (see Power of χ^2_{tdt} and χ^2_{asp}) except that an adjusted significance cutpoint would be used as described above.

To briefly examine power for a particular multi-allelic example, consider the bi-allelic marker and disease locus in the bottom line of Table 2 ($r = 4$). The frequencies (p and m) of disease allele D and positively associated marker allele A are 0.15 and 0.25, respectively, and TDT power is 0.99 when $\delta = \delta_{max}$ and 0.86 when $\delta = \frac{1}{2}\delta_{max}$. Suppose p and m remain constant as does the degree of positive association between alleles A and D ($\delta = \delta_{max}$ or $\frac{1}{2}\delta_{max}$) but suppose the marker consists of $k-1$ additional (non-A) alleles having negative or no association with disease allele D. Then A/non-A would give the highest $E(\chi^2_{tdt})$ of any i/non-i and thus TDT power would be determined by the P_t and H/F shown in the bottom line of the table. According to Ewens & Spielman (1997), adjusted cutpoints (0.05 significance) for $k = 2$, $k = 4$ and $k = 8$ are $\chi^2_{tdt} = 3.84$, 6.10 and 7.41, respectively; thus TDT power when $k = 2$, 4 or 8 would be 0.86, 0.71 or 0.61 at $\delta = \frac{1}{2}\delta_{max}$ and would be 0.99 for $k \leqslant 8$ when $\delta = \delta_{max}$. This example suggests that TDT power for a multi-allelic marker remains relatively strong if (a) one marker allele is strongly associated with either allele of a bi-allelic disease locus and (b) the two associated alleles have similar population frequencies.

Concluding remarks

Strength of evidence for linkage provided by χ^2_{asp} and χ^2_{tdt} critically depends upon the magnitude of departure from the null hypothesis value of 0.5, the size of departure being quantified by $(P_s-0.5)$ and $|P_t-0.5|$ for the ASP and TDT paradigms, respectively. In this paper, I have shown that $(P_s-0.5)$ and $|P_t-0.5|$ are each a product of three corresponding factors [$(P_s-0.5) = L_s M_s R_s$, $|P_t-0.5| = L_t|M_t|R_t$]. L_s and L_t depend only on the recombination fraction (θ), R_s and R_t depend only on disease penetrance (α, β, γ) and the frequency (p) of the disease allele and, furthermore, marker allele frequency (m) and disequilibrium (δ) influence only M_s and $|M_t|$. Hence, the corresponding factors in $(P_s-0.5)$ and $|P_t-0.5|$ facilitate comparisons between the ASP and TDT paradigms, and also enable some 'partitioning' of the contribution to evidence for linkage provided by standard genetic variables such as θ, δ, m, etc.

Together with the expression for parental heterozygosity at the marker (H/F), the expressions for P_s and P_t provide a general framework for calculating and comparing the power of χ^2_{asp} and χ^2_{tdt}. This framework generalizes the ASP-TDT comparison of Risch & Merikangas (1996) by encompassing many modes of inheritance rather than just one, and also by enabling TDT power to be [325.12]

[325-12]

calculated for a marker that is distinct from the disease locus. Analysis of the equations shows that TDT power is greatly increased if disequilibrium is strong and if the disease allele and positively associated marker allele have similar population frequencies. The equations also show that the superior power of the TDT compared to the ASP test is greatest when susceptibility loci confer modest disease risk, as indicated by low values of the penetrance ratio r. When a marker is strongly associated with a disease locus that contributes modest disease risk, $|P_t - 0.5| \gg (P_s - 0.5) \approx 0$. Thus, the TDT is likely to play an important role in detecting and replicating linkages to loci responsible for complex genetic disease.

I am deeply grateful to Richard Spielman for encouragement and valuable suggestions as this work developed. I am also indebted to Warren Ewens for valuable comments and for criticism that improved the manuscript. This research was supported by NIH grants DK46618 and DK47481 and by grant 193189 from the Juvenile Diabetes Foundation.

REFERENCES

Bell, G. I., Horita, S. & Karam, J. H. (1984) A polymorphic locus near the human insulin gene is associated with insulin-dependent diabetes mellitus. *Diabetes* 33, 176–183.

Blackwelder, W. C. & Elston, R. C. (1985) A comparison of sib-pair linkage tests for susceptibility loci. *Genet. Epidemiol.* 2, 85–97.

Clerget-Darpoux, F., Babron, M. C. & Bickeböller, H. (1996) Comparing the power of linkage detection by the transmission disequilibrium test and identity-by-descent test. *Genet. Epidemiol.* 12, 583–588.

Cox, N. J. & Spielman, R. S. (1989) The insulin gene and susceptibility to IDDM. *Genet Epidemiol* 6, 65–69.

Ewens, W. J. & Spielman, R. S. (1997) Disease associations and the transmission/disequilibrium test (TDT). *Current Protocols in Human Genetics* 1.12.1–1.12.13.

Julier, C., Hyer, R. N., Davies, J., Merlin, F., Soularue, P., Briant, L., Cathelineau, G., *et al.* (1991) Insulin-IGF2 region on chromosome 11p encodes a gene implicated in HLA-DR4–dependent diabetes susceptibility. *Nature* 354, 155–159.

Kaplan, N. L., Martin, E. R. & Weir, B. S. (1997) Power studies of the transmission/disequilibrium tests with multiple alleles. *Am. J. Hum. Genet.* 60, 691–702.

McGinnis, R. E., Spielman, R. S. & Ewens, W. J. (1991) Linkage between the insulin gene (IG) region and susceptibility to insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet. Suppl.* 49, A476.

Müller-Myhsok, B. & Abel, L. (1997) Genetic analysis of complex diseases. *Science* 275, 1328–1329.

Ott, J. (1989) Statistical properties of the haplotype relative risk. *Genet. Epidemiol.* 6, 127–130.

Ott, J. (ed) (1991) *Analysis of human genetic linkage.* Johns Hopkins University Press, Baltimore.

[325.08]

Risch, N. (1990) Linkage strategies for genetically complex traits. I. Multilocus models. *Am. J. Hum. Genet.* 46, 229–241.

Risch, N. & Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517.

Pearson, E. S. & Hartley, H. O. (ed) (1954) *Biometrika tables for statisticians.* Vol 1. Cambridge University Press.

Sorajd, D. J. & Sommer, S. S. (1994) Comparison of statistics for candidate-gene association studies using cases and parents. *Am. J. Hum. Genet.* 55, 402–409.

Sham, P. C. & Curtis, D. (1995) An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann. Hum. Genet.* 59, 323–336.

Spielman, R. S., Baur, M. P. & Clerget-Darpoux, F. (1989). Genetic analysis of IDDM: summary of GAW5 IDDM results. *Genet. Epidemiol.* 6, 43–58.

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52, 506–516.

Spielman, R. S. & Ewens, W. J. (1996) The TDT and other family-based tests for linkage disequilibrium and association. *Am. J. Hum. Genet.* 59, 983–989.

Terwilliger, J. D. & Ott, J. (1992) A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum. Hered.* 42, 337–346.

[325.08]

Thomson, G., Robinson, W. P., Kuhner, M. K. & Joe, S. (1989) HLA, insulin gene, and Gm associations with IDDM. *Genet. Epidemiol.* 6, 155–160.

Weir, B. S. (ed) (1996) *Genetic data analysis II.* 2nd ed., Sinauer Associates Inc., Sunderland, MA.

APPENDIX I

Derivation of expressions for P_s, P_t, H

The derivations assume the general model of a bi-allelic marker and linked bi-allelic disease locus that is the only locus that underlies disease susceptibility (see General algebraic model of linkage in the main text). I begin the derivation of P_s and P_t (equations (1) and (2)) by first deriving

mating type in Table A2, this sum simplifies to the overall 'weighted' subpopulation frequency shown in column 2. For example, when $k=2$ (and hence $f_w = r(r-1)/(N-1)$), the sum for the mating type in line 1 simplifies as follows:

$$\sum_{r=2}^{N} r\frac{r-1}{N-1}2(c_1c_4)p^2\binom{N}{r}\left(\frac{\alpha+\beta}{2}\right)^r\left(1-\frac{\alpha+\beta}{2}\right)^{N-r}$$

$$= 2(c_1c_4)p^2N\left(\frac{\alpha+\beta}{2}\right)^2\sum_{r=2}^{N}\binom{N-2}{r-2}\left(\frac{\alpha+\beta}{2}\right)^{r-2}\left(1-\frac{\alpha+\beta}{2}\right)^{N-r}$$

$$= 2(c_1c_4)p^2N\left(\frac{\alpha+\beta}{2}\right)^2\sum_{r=0}^{N-2}\binom{N-2}{r}\left(\frac{\alpha+\beta}{2}\right)^{r}\left(1-\frac{\alpha+\beta}{2}\right)^{N-2-r}$$

$$= 2(c_1c_4)p^2N\left(\frac{\alpha+\beta}{2}\right)^2\left(\frac{\alpha+\beta}{2}+1-\frac{\alpha+\beta}{2}\right)^{N-2}$$

$$= 2(c_1c_4)p^2N\left(\frac{\alpha+\beta}{2}\right)^2.$$

Based on the 'weighted' subpopulation frequencies in column 2, what then is the probability of randomly selecting a particular 'A/B father-mating type' from among families with an A/B father and N children, at least two of whom are affected? Setting $k=2$, the probability (P_{rs}) of randomly selecting a particular mating type would be the weighted frequency of the mating type divided by the sum of all the weighted frequencies in column 2. Thus, the probability of randomly selecting the mating type on line 1 ($AD/Bd \times D/D$) would be:

$$P_{rs} = \frac{2(c_1c_4)}{H}p^2\left(\frac{\alpha+\beta}{2}\right)^2,$$

where
$$H = 2(c_1c_4+c_2c_3)\,p^2\left(\frac{\alpha+\beta}{2}\right)^2+\tfrac{1}{2}p(1-p)\left(\frac{\alpha+2\beta+\gamma}{2}\right)^2+(1-p)^2\left(\frac{\beta+\gamma}{2}\right)^2$$
$$+2c_1c_3\{p^2\alpha^2+\tfrac{1}{2}p(1-p)(\alpha+\beta)^2+(1-p)^2\beta^2\}$$
$$+2c_2c_4\{p^2\beta^2+\tfrac{1}{2}p(1-p)(\beta+\gamma)^2+(1-p)^2\gamma^2\}.$$

[0325.14]

Note that the quantity N cancels in the numerator and denominator of P_{rs}, thus demonstrating that the probability is independent of family size (N) and hence applies to mixed populations of families of *any* size that have 2 or more affected offspring. Furthermore, if mating types with A/B mothers were included in Table A2, the only effect would be to double each frequency in column 2; but P_{rs} would be unchanged since the '2's' in the doubled numerator and denominator of P_{rs} would cancel. Thus, by setting $k=2$, the P_{rs} calculated for each mating type in Table A2 applies to random selection of A/B *parents* of two or more affected offspring from families of *any* size.

What, then, is the probability that a randomly selected A/B parent of a particular mating type transmitted allele A (allele B) to an affected child? For each mating type in Table A2, the two rightmost columns show the conditional probability that the A/B parent transmitted allele A or B to an individual affected offspring. These probabilities follow directly from the conditional probabilities in Table A1. For example, in the $AD/Bd \times D/D$ mating in line 1 of Table A2, the A/B parent has allele A in coupling with allele D, and B in coupling with d, and thus Table A1 implies that A is transmitted to affected offspring with a probability of

$$(1-\theta)\frac{\alpha}{\alpha+\beta}+\theta\frac{\beta}{\alpha+\beta}=\frac{\alpha-\theta(\alpha-\beta)}{\alpha+\beta}.$$